Restrained Revision

Richard Booth
School of IT and Computer Science
University of Wollongong
Wollongong, NSW 2522, Australia

Samir Chopra
Department of Computer Science
Brooklyn College of CUNY
Brooklyn, NY 11210

Thomas Meyer
National ICT Australia and
School of Computer Science and Engineering
University of New South Wales
Sydney, NSW 2052 Australia

Abstract

As part of the justification of their proposed framework for iterated belief revision Darwiche and Pearl advanced a convincing argument against Boutilier's natural revision, and provided a prototypical revision operator which fits into their scheme. We show that the Darwiche-Pearl arguments lead naturally to the acceptance of a smaller class of operators which we refer to as *admissible*. These are characterised in terms of syntactic as well as semantic postulates. Admissible revision ensures that the penultimate input is not ignored completely, thereby eliminating natural revision, but includes the prototypical Darwiche-Pearl operator, Nayak's well-known lexicographic revision operator, and a newly introduced operator called *restrained revision*. We give a syntactic and a semantic characterisation of restrained revision, and demonstrate that it satisfies desirable properties. In particular, we show that it is the most conservative of admissible revision operators, while lexicographic revision is the least conservative. This makes an interesting comparison with the Darwiche-Pearl framework in which lexicographic revision is also the least conservative, but natural revision is the most conservative. In a sense, then, restrained revision can be seen as an appropriate replacement for natural revision. Finally we show that restrained revision can also be viewed as a composite operator, consisting of natural revision preceded by an application of a "backwards revision" operator previously studied by Papini.

1 Introduction

Many formal treatments of iterated belief revision rely on recipes for manipulating plausibility orderings, usually total preorders, of possible worlds. Typically, these require that (a) the epistemic input α is entailed by the knowledge base associated with the revised plausibility ordering, and (b) that the remaining worlds are arranged in some plausible ordering corresponding to a rational change of an agent's beliefs. This ensures that (a) we can obtain a new knowledge base from the lowest rank in the plausibility ordering of possible worlds and (b) that a new ranking of possible worlds is available as a revision target for the next epistemic input.

Most iterated revision schemes are sensitive to the history of belief changes[1], based on a version of the 'most recent is best' argument, where the newest information is of higher priority than anything else in the knowledge base. Arguably the most extreme case of this is Nayak's lexicographic revision [11, 12]. However, there are operators where, once *admitted* to the knowledge base, it rapidly becomes as much of a candidate for removal as anything else in the set when another, newer, piece of information comes along, Boutilier's natural revision [4, 5] being a case in point. (A dual to this is what Rott[14] terms *radical* revision where the new information is accepted with maximal, irremediable entrenchment – see also [16]). Another issue to consider is the problem termed *temporal incoherence* by Rott [14]:

[1] An *external* revision scheme like [2, 7] is not.

the comparative recency of information should translate systematically into comparative importance, strength or entrenchment

In an influential paper Darwiche and Pearl [6] propose a framework for iterated revision. Their proposal is characterised in terms of sets of syntactic and semantic postulates, but can also be viewed from the perspective of conditional beliefs. To justify their proposal Darwiche and Pearl mount a comprehensive argument. The argument includes a critique of natural revision, which is shown to admit too few changes. In addition, they provide a concrete revision operator which is shown to satisfy their postulates. In many ways this can be seen as the prototypical Darwiche-Pearl operator. It is instructive to observe that the two best-known operators satisfying the Darwiche-Pearl postulates, natural revision and lexicographic revision, form the opposite extremes of the Darwiche-Pearl framework: Natural revision is the most conservative Darwiche-Pearl operator, while lexicographic revision is the least conservative of the Darwiche-Pearl operators.

In this paper we show that the Darwiche-Pearl arguments lead naturally to the acceptance of a smaller class of operators which we refer to as *admissible*. We provide characterisations of admissible revision, in terms of syntactic as well as semantic postulates. Admissible revision ensures that the penultimate input is not ignored completely. A consequence of this is that natural revision is eliminated. On the other hand, admissible revision includes the prototypical Darwiche-Pearl operator as well as lexicographic revision, the latter result also showing that lexicographic revision is the least conservative of the admissible operators. The removal of natural revision from the scene leaves a gap which is filled by the introduction of a new operator we refer to as *restrained revision*. It is the most conservative of admissible revision operators, and can thus be seen as an appropriate replacement of natural revision. We give a syntactic and a semantic characterisation of restrained revision, and demonstrate that it satisfies desirable properties. In particular, and unlike lexicographic revision, it ensures that older information is not discarded unnecessarily, and it shows that the problem of temporal incoherence can be dealt with.

Although natural revision does not feature in the class of admissible revision operators, we nevertheless show that it still has a role to play in iterated revision, provided it is first tempered appropriately. We show that restrained revision can also be viewed as a composite operator, consisting of natural revision preceded by an application of a "backwards revision" operator previously studied by Papini [13].

The paper is organised as follows. After outlining some notation, we review the Darwiche-Pearl framework in Section 2. This is followed by a discussion of admissible revision in Section 3. In Section 4 we introduce restrained revision, and in Section 5 we show how it can be defined as a composite operator. Section 6 concludes and briefly discusses some future work.

1.1 Notation

We assume a finitely generated propositional language L which includes the constants $\top$ and $\bot$, is closed under the usual propositional connectives, and is equipped with a classical model-theoretic semantics. V is the set of valuations of L and $[\alpha]$ (or $[B]$) is the set of models of $\alpha \in L$ (or $B \subseteq L$). Classical entailment is denoted by $\models$ and logical equivalence by $\equiv$. Greek letters $\alpha, \beta, \ldots$ stand for arbitrary formulas.

2 Darwiche-Pearl Revision

In this section we briefly review the Darwiche and Pearl [6] approach to iterated belief revision. Darwiche and Pearl reformulated the AGM postulates [1] to be compatible with their suggested approach to iterated revision. This necessitated a move from knowledge bases to *epistemic states*. An epistemic state contains, in addition to a knowledge base, all the information needed for coherent reasoning including, in particular, the strategy for belief revision which the agent wishes to employ at a given time. This includes a plausibility ordering on all valuations, a total preorder, with elements lower down in the ordering deemed more plausible.

Definition 1 *From every epistemic state $\mathbb{E}$ can be extracted a total preorder on valuations $\preceq_{\mathbb{E}}$, and a consistent knowledge base $B(\mathbb{E})$.*[2] *$\min(\alpha, \preceq_{\mathbb{E}})$ denotes the minimal models of α under $\preceq_{\mathbb{E}}$. The knowledge base associated with the epistemic state is obtained by considering the minimal models in $\preceq_{\mathbb{E}}$ i.e., $[B(\mathbb{E})] = \min(\top, \preceq_{\mathbb{E}})$.*

[2]The requirement that $B(\mathbb{E})$ be consistent enables us to obtain a unique knowledge base from the total preorder $\preceq_{\mathbb{E}}$. Preservation of the results in this paper when this requirement is relaxed is possible, but technically messy.

In the reformulated postulates $*$ is a belief change operator on epistemic states, not knowledge bases.

($\mathbb{E}$*1) $B(\mathbb{E} * \alpha) = Cn(B(\mathbb{E} * \alpha))$

($\mathbb{E}$*2) $\alpha \in B(\mathbb{E} * \alpha)$

($\mathbb{E}$*3) $B(\mathbb{E} * \alpha) \subseteq B(\mathbb{E}) + \alpha$

($\mathbb{E}$*4) If $\neg\alpha \notin B(\mathbb{E})$ then $B(\mathbb{E}) + \alpha \subseteq B(\mathbb{E} * \alpha)$

($\mathbb{E}$*5) If $\alpha \equiv \beta$ then $\mathbb{E} * \alpha = \mathbb{E} * \beta$

($\mathbb{E}$*6) $\bot \in B(\mathbb{E} * \alpha)$ iff $\models \neg\alpha$

($\mathbb{E}$*7) $B(\mathbb{E} * (\alpha \wedge \beta)) \subseteq B(\mathbb{E} * \alpha) + \beta$

($\mathbb{E}$*8) If $\neg\beta \notin B(\mathbb{E} * \alpha)$ then $B(\mathbb{E} * \alpha) + \beta \subseteq B(\mathbb{E} * (\alpha \wedge \beta))$

The observant reader will note that our assumption of a consistent $B(\mathbb{E})$ is incompatible with a successful revision by $\bot$. This requires that we jettison ($\mathbb{E}$*6) and insist on consistent epistemic inputs only.[3] We shall refer to the reformulated AGM postulates, with ($\mathbb{E}$*6) removed, as RAGM. The main reason for the reformulation occurs in ($\mathbb{E}$*5), which states that revising by logically equivalent formulas results in the same epistemic state. The original AGM postulate requires only that the knowledge base extracted from the resulting epistemic state be the same after revision by logically equivalent formulas.

RAGM guarantees a unique extracted knowledge base (modulo logical equivalence) when revision by α is performed. It sets $[B(\mathbb{E} * \alpha)]$ equal to $\min(\alpha, \preceq_{\mathbb{E}})$ and thereby fixes the most plausible valuations in $\preceq_{\mathbb{E}*\alpha}$. What is not fixed is how to order the remaining valuations.

We now list the Darwiche-Pearl postulates for iterated revision [6].

(C1) If $\alpha \models \beta$ then $B(\mathbb{E} * \beta * \alpha) = B(\mathbb{E} * \alpha)$

(C2) If $\alpha \models \neg\beta$ then $B(\mathbb{E} * \beta * \alpha) = B(\mathbb{E} * \alpha)$

(C3) If $\beta \in B(\mathbb{E} * \alpha)$ then $\beta \in B(\mathbb{E} * \beta * \alpha)$

(C4) If $\neg\beta \notin B(\mathbb{E} * \alpha)$ then $\neg\beta \notin B(\mathbb{E} * \beta * \alpha)$

The following are the corresponding semantic versions (with $v, w \in V$):

(CR1) If $v \in [\alpha], w \in [\alpha]$ then $v \preceq_{\mathbb{E}} w$ iff $v \preceq_{\mathbb{E}*\alpha} w$

(CR2) If $v \in [\neg\alpha], w \in [\neg\alpha]$ then $v \preceq_{\mathbb{E}} w$ iff $v \preceq_{\mathbb{E}*\alpha} w$

(CR3) If $v \in [\alpha], w \in [\neg\alpha]$ then $v \prec_{\mathbb{E}} w$ only if $v \prec_{\mathbb{E}*\alpha} w$

(CR4) If $v \in [\alpha], w \in [\neg\alpha]$ then $v \preceq_{\mathbb{E}} w$ only if $v \preceq_{\mathbb{E}*\alpha} w$

Darwiche and Pearl showed that, given RAGM, a precise correspondence obtains between (Ci) and (CRi) above ($i = 1, 2, 3, 4$). The postulate (C1) states that when two pieces of information—one more specific than the other—arrive, the first is made redundant by the second. (C2) says that when two contradictory epistemic inputs arrive, the second one prevails; the second evidence alone yields the same knowledge base. (C3) says that a piece of evidence β should be retained after accommodating more recent evidence α that entails β given the current knowledge base. (C4) simply says that no epistemic input can act as its own defeater. We shall refer to the belief revision operators satisfying RAGM and (C1) to (C4) as *DP-operators*.

One of the guiding principles of belief revision is the principle of minimal change: changes to a belief state ought to be kept to a minimum. What is not always clear is what ought to be minimised. In AGM theory the prevailing wisdom is that minimal change refers to the sets of sentences corresponding to knowledge bases. But there are other interpretations. With the move from knowledge bases to epistemic states, minimal change can be defined in terms of the fewest possible changes to the associated plausibility ordering $\preceq_{\mathbb{E}}$. In what follows we shall frequently have the opportunity to refer to the latter interpretation of minimal change.

[3] The part of ($\mathbb{E}$*6) which requires a consistent $B(\mathbb{E} * \alpha)$ is rendered superfluous by ($\mathbb{E}$*1) and the assumption that knowledge bases extracted from all epistemic states have to be consistent.

3 Admissible Revision

In this section we consider two of the best-known DP-operators, and propose a strengthening of the Darwiche-Pearl framework. This strengthening is suggested by some of the arguments advanced by Darwiche and Pearl themselves. The strengthening eliminates one of the operators they criticise, and is satisfied by the sole operator they provide as an instance of their framework.

The oldest known DP-operator is Boutilier's *natural revision* [4, 5]. Its main feature is the application of the principle of minimal change to epistemic states. It is characterised by RAGM plus the following postulate:

(CB) If $\neg\beta \in B(\mathbb{E} * \alpha)$ then $B(\mathbb{E} * \alpha * \beta) = B(\mathbb{E} * \beta)$

(CB) requires that, whenever $B(\mathbb{E} * \alpha)$ is inconsistent with β, revising $\mathbb{E} * \alpha$ with β will completely ignore the revision by α. Its semantic counterpart is as follows:

(CBR) For $v, w \notin [B(\mathbb{E} * \alpha)]$, $v \preceq_{\mathbb{E}*\alpha} w$ iff $v \preceq_{\mathbb{E}} w$

From (CBR) it is clear that natural revision is an application of minimal change to epistemic states. It requires that, barring the changes mandated by RAGM, the relative ordering of valuations remains unchanged. So natural revision is the most conservative of *all* DP-operators. Such a strict adherence to minimal change is inadvisable and needs to be tempered appropriately, an issue that will be addressed in Section 5. Darwiche and Pearl have shown that (CB) is too strong, and that natural revision is not all that natural, sometimes yielding counterintuitive results.

Example 1 *[6] We encounter a strange animal and it appears to be a bird, so we believe it is one. As it comes closer, we see clearly that the animal is red, so we believe it is a red bird. To remove further doubts we call in a bird expert who examines it and concludes that it is not a bird, but some sort of animal. Should we still believe the animal is red? (CB) tells us we should no longer believe it is red. This can be seen by substituting $B(\mathbb{E}) = Cn(\neg\beta) = Cn(bird)$ and $\alpha \equiv red$ in (CB), instructing us to totally ignore the observation α as if it had never taken place.*

Given this example it is perhaps surprising that Darwiche and Pearl never considered the following postulate:

(P) If $\neg\beta \notin B(\mathbb{E} * \alpha)$ then $\beta \in B(\mathbb{E} * \beta * \alpha)$

Recently (P) was also proposed in [9] where it is referred to as a postulate of *Independence*. It states that whenever β is consistent with a revision by α, it should be retained if a β-revision is inserted just before the α-revision. Applying this to Example 1 we see that, since *red* is consistent with $B(\mathbb{E} * \neg bird)$, we have $red \in B(\mathbb{E} * red * \neg bird)$; that is, we have to believe the animal, which we now know not to be a bird, is red. The semantic counterpart of (P) looks like this:

(PR) For $v \in [\alpha]$ and $w \in [\neg\alpha]$, if $v \preceq_{\mathbb{E}} w$ then $v \prec_{\mathbb{E}*\alpha} w$

(PR) requires an α-world v that is at least as plausible as a $\neg\alpha$-world w to be strictly more plausible than w after an α-revision.

Proposition 1 *If $*$ satisfies RAGM, then it satisfies (P) iff it also satisfies (PR).*

Note that (P) has the antecedent of (C4) and the consequent of (C3). Thus it follows that (P) is stronger than (C3) and (C4) combined. This is easily seen from the semantic counterparts of these postulates as well. It also follows that the only concrete example of an iterated revision operator provided by Darwiche and Pearl (the operator they refer to as $\bullet$ and which employs a form of Spohnian conditioning [17]) satisfies (PR), and therefore (P) as well. Furthermore, by adopting (P) we explicitly exclude natural revision as a permissible operator. So accepting (P) is a move towards the viewpoint that information obtained before the latest input ought not to be discarded unnecessarily.

Based on this analysis we propose a strengthening of the Darwiche-Pearl framework in which (C3) and (C4) are replaced by (P).

Definition 2 *A revision operator is* admissible *iff it satisfies RAGM, (C1), (C2), and (P).*

Inasmuch as the Darwiche-Pearl framework can be visualised semantically as one in which α-worlds slide "downwards" relative to $\neg\alpha$-worlds, admissible revision ensures that this "downwards" slide is a strict one.

Another view of (P) is that it is a significant weakening of the following property, first introduced in [12]:

(Recalcitrance) If $\neg\beta \notin Cn(\alpha)$ then $\beta \in B(\mathbb{E} * \beta * \alpha)$

Semantically, (Recalcitrance) corresponds to the following property, as was pointed out by Booth in [3] and implicitly contained in [12]:

(R) For $v \in [\alpha]$, $w \in [\neg\alpha]$, $v \prec_{\mathbb{E}*\alpha} w$

(Recalcitrance) is a property of Nayak's *lexicographic revision* operator [11, 12], the second of the well-known DP-operators we consider. In fact, lexicographic revision is characterised by RAGM together with (C1), (C2) and (Recalcitrance), a result that is easily proved from the semantic counterparts of these properties and Nayak et al.'s semantic characterisation of lexicographic revision in [12]. An analysis of the semantic characterisation of lexicographic revision shows that it is the *least* conservative of the DP-operators, effecting the most changes in the relative ordering of valuations permitted by RAGM and the Darwiche-Pearl postulates. Since it is also an admissible revision operator, it follows that it is also the least conservative admissible operator.

The problem with (Recalcitrance) is that the decision of whether to accept β after a subsequent revision by α is completely determined by the logical relationship between β and α – the epistemic state $\mathbb{E}$ is robbed of all influence. The replacement of (Recalcitrance) by the weaker (P) already gives $\mathbb{E}$ more influence in the outcome. In the next section we constrain matters further by giving $\mathbb{E}$ as much influence as allowed by the postulates for admissible revision. Such a move ensures greater sensitivity to the agent's epistemic record in making further changes.

Note that lexicographic revision assumes that more recent information takes complete precedence over information obtained previously. Thus, when applied to Example 1, it requires us to believe that the animal, which we have previously assumed to be a bird, is indeed red, primarily because *red* is a recent input which does not conflict with the most recently obtained input. While this is a reasonable approach in many circumstances, a dogmatic adherence to it can be problematic, as the following example shows.

Example 2 *We observe a creature which is clearly red, but we are too far away to determine whether it is a bird or an animal. So we adopt the knowledge base $B(\mathbb{E}) \equiv red$. Next to us is a person who declares that, since the creature is red, it must be a bird. We have no reason to doubt him, and so we adopt the belief $red \rightarrow bird$. Now the creature moves closer and it becomes clear that it is not a bird. The question is, should we continue believing that it is red? Under the circumstances described above we want our initial observation to take precedence over the opinion of the self-proclaimed expert and believe that the animal is red. But lexicographic revision does not allow us to do so.*

While (P) allows for the possibility of retaining the belief that the animal is red, it does not *enforce* this belief. Below we provide a property which does so.[4] To help us express this property, we introduce an extra piece of terminology and notation:

Definition 3 *α and β* counteract *with respect to an epistemic state $\mathbb{E}$, written $\alpha \leftrightsquigarrow_{\mathbb{E}} \beta$, iff $\neg\beta \in B(\mathbb{E} * \alpha)$ and $\neg\alpha \in B(\mathbb{E} * \beta)$.*

The use of the term *counteract* to describe this relation is taken from [12]. $\alpha \leftrightsquigarrow_{\mathbb{E}} \beta$ means that, from the viewpoint of $\mathbb{E}$, α and β tend to "exclude" each other. Some points to note about $\leftrightsquigarrow_{\mathbb{E}}$ are (a) it is symmetric, and (b) it depends only on the total preorder $\preceq_{\mathbb{E}}$ obtained from $\mathbb{E}$. Furthermore, if α and β are logically inconsistent then $\alpha \leftrightsquigarrow_{\mathbb{E}} \beta$, but the converse need not hold. Thus $\leftrightsquigarrow_{\mathbb{E}}$ can be seen as a weak form of inconsistency. Now consider the following property:

(D) If $\alpha \leftrightsquigarrow_{\mathbb{E}} \beta$ then $\neg\alpha \in B(\mathbb{E} * \alpha * \beta)$

[4] Of course, the order of the sentences in the revision sequence is important, and changing it will have an effect on the outcome.

(D) requires that, whenever α and β counteract with respect to $\mathbb{E}$, α should be *disallowed* when an α-revision is followed by a β-revision. In other words, when the β-revision of $\mathbb{E} * \alpha$ takes place, the information encoded in $\mathbb{E}$ takes precedence over the information contained in $\mathbb{E} * \alpha$. Darwiche and Pearl considered this property (it is their rule (C6) in [6]), but argued against it, citing the following example.

Example 3 *[6] We believe that exactly one of John and Mary committed a murder. Now we get persuasive evidence indicating that John is the murderer. This is followed by persuasive information indicating that Mary is the murderer. Let α represent that John committed the murder and β that Mary committed the murder. Then (D) forces us to conclude that Mary, but not John, was involved in the murder. This, according to Darwiche and Pearl, is counterintuitive, since we should conclude that both were involved in committing the murder.*

Darwiche and Pearl's argument against (D) rests upon the assumption that more recent information ought to take precedence over information previously obtained. But as we have seen in Example 2, this is not always a valid assumption. In fact, the application of (D) to Example 2, with $\alpha = red \rightarrow bird$ and $\beta = \neg bird$, produces the intuitively correct result of a belief in the observed animal being red: $red \in B(\mathbb{E} * (red \rightarrow bird) * \neg bird)$.

Another way to gain insight into the significance of (D) is to consider its semantic counterpart:

(DR) For $v \in [\neg\alpha]$, $w \in [\alpha]$, and $w \notin [B(\mathbb{E} * \alpha)]$, if $v \prec_{\mathbb{E}} w$ then $v \prec_{\mathbb{E}*\alpha} w$

(DR) curtails the rise in plausiblity of α-worlds after an α-revision. It ensures that, with the exception of the most plausible α-worlds, the relative ordering between an α-world and the $\neg\alpha$-worlds more plausible than it remains unchanged.

Proposition 2 *Whenever a revision operator $*$ satisfies RAGM, then $*$ satisfies (D) iff it satisfies (DR).*

4 Restrained Revision

In this section we strengthen the requirements on admissible revision (those operators satisfying RAGM, (C1), (C2) and (P)) by insisting that (D) is satisfied as well. To do so, let us first consider the semantic definition of an interesting admissible revision operator. Recall that RAGM fixes the set of $(\preceq_{\mathbb{E}*\alpha})$-minimal models, setting them equal to $\min(\alpha, \preceq_{\mathbb{E}})$, but places no restriction on how the remaining valuations should be ordered. The following property provides a *unique* relative ordering of the remaining valuations.

(R) $\forall v, w \notin [B(\mathbb{E} * \alpha)]$, $v \preceq_{\mathbb{E}*\alpha} w$ iff $\begin{cases} v \prec_{\mathbb{E}} w \text{ or,} \\ v \preceq_{\mathbb{E}} w \text{ and } (v \in [\alpha] \text{ or } w \in [\neg\alpha]) \end{cases}$

(R) says that the relative ordering of the valuations that are not $(\preceq_{\mathbb{E}*\alpha})$-minimal remains unchanged, except for α-worlds and $\neg\alpha$-worlds on the same plausibility level; those are split into two levels with the α-worlds more plausible than the $\neg\alpha$-worlds. So RAGM combined with (R) fixes a unique operator.

Definition 4 *The (unique) revision operator satisfying RAGM and (R) is called* restrained revision.

It turns out that restrained revision is the *only* admissible revision operator satisfying (D).

Theorem 1 *RAGM, (C1), (C2), (P) and (D) provide an exact characterisation of restrained revision.*

The proof is easily obtained from the semantic counterparts of these properties.

Another interpretation of (R) is that it maintains the relative ordering of the valuations that are not $(\preceq_{\mathbb{E}*\alpha})$-minimal, except for the changes mandated by (PR). From this it can be seen that restrained revision is the most conservative of all admissible revision operators. So, in the context of admissible revision, restrained revision takes on the role played by natural revision in the Darwiche-Pearl framework.

Another look at Examples 2 and 3 shows that they share some interesting structural properties. In both examples the initial knowledge base $B(\mathbb{E})$ is pairwise consistent with each of the subsequent sentences in the revision sequence, while the sentences in each revision sequence are pairwise inconsistent. And in both examples the information contained in the initial knowledge base $B(\mathbb{E})$ is retained after the revision sequence. These commonalities are instances of an important general result. Let Γ denote the non-empty sequence of inputs $\gamma_1, \ldots, \gamma_n$, and let $\mathbb{E} * \Gamma$ denote the revision sequence $\mathbb{E} * \gamma_1 * \ldots * \gamma_n$. Furthermore we shall refer to an epistemic state $\mathbb{E}$ as Γ-*compatible* provided that $\neg\gamma_i \notin B(\mathbb{E})$ for every i in $\{1, \ldots, n\}$.

(O) If $\mathbb{E}$ is Γ-compatible then $B(\mathbb{E}) \subseteq B(\mathbb{E} * \Gamma)$

(O) says that as long as $B(\mathbb{E})$ is not in direct conflict with *any* of the inputs in the sequence $\gamma_1, \ldots, \gamma_n$, the entire $B(\mathbb{E})$ *has* to be propagated to the knowledge base obtained from the revision sequence $\mathbb{E} * \gamma_1 * \ldots * \gamma_n$. This is a very desirable preservation property, and one that is satisfied by restrained revision.

Proposition 3 *Restrained revision satisfies (O).*

Although restrained revision preserves information which has not been directly contradicted, it is not dogmatically wedded to older information. If neither of two successive, but incompatible, epistemic states are in conflict with any of the inputs of a sequence $\Gamma = \gamma_1, \ldots, \gamma_n$, it prefers the latter epistemic state when revising by Γ.

Proposition 4 *Restrained revision satisfies the following property:*

(Q) *If $\mathbb{E}$ and $\mathbb{E} * \alpha$ are both Γ-compatible but $B(\mathbb{E}) \cup B(\mathbb{E} * \alpha) \models \bot$, then $B(\mathbb{E} * \alpha) \subseteq B(\mathbb{E} * \alpha * \Gamma)$ and $B(\mathbb{E}) \nsubseteq B(\mathbb{E} * \alpha * \Gamma)$*

Next we consider another preservation property, but this time, unlike the case for (O) and (Q), we look at circumstances where $B(\mathbb{E})$ is incompatible with some of the inputs in a revision sequence.

(S) If $\neg\beta \in B(\mathbb{E} * \alpha)$ and $\neg\beta \in B(\mathbb{E} * \neg\alpha)$ then $B(\mathbb{E} * \alpha * \neg\alpha * \beta) = B(\mathbb{E} * \alpha * \beta)$

Note that, given RAGM, the antecedent of (S) implies that $\neg\beta \in B(\mathbb{E})$. Thus (S) states that if $\neg\beta$ is believed initially, and that a subsequent commitment to either α or its negation would not change this fact, then after the sequence of inputs in which β is preceded by α and $\neg\alpha$, the *second* input concerning α is nullified, and the older input regarding α is retained.

Proposition 5 *Restrained revision satisfies (S).*

We now turn to two properties first mentioned (as far as we know) by Schlechta et al. in [15] (see also [10]):

(Disj1) $B(\mathbb{E} * \alpha * \beta) \cap B(\mathbb{E} * \gamma * \beta) \subseteq B(\mathbb{E} * (\alpha \vee \gamma) * \beta)$

(Disj2) $B(\mathbb{E} * (\alpha \vee \gamma) * \beta) \subseteq B(\mathbb{E} * \alpha * \beta) \cup B(\mathbb{E} * \gamma * \beta)$

(Disj1) says that if a sentence is believed after any one of two sequences of revisions that differ only at step i (step i being α in one case and γ in the other), then the sentence should also be believed after that sequence which differs from both only in that step i is a revision by the disjunction $\alpha \vee \gamma$. Similarly, (Disj2) says that every sentence believed after an $(\alpha \vee \gamma)$-β-revision should be believed after at least one of (α-β) and (γ-β). Both condititions are reasonable properties to expect of revision operators.

Proposition 6 *Restrained revision satisfies (Disj1) and (Disj2).*

It can be shown that lexicographic revision also satisfies these two rules.

To conclude this section we show that there is a more compact syntactic representation of restrained revision. First we show that (C1) and (P) can be combined into a single property, and so can (C2) and (D).

Proposition 7 *Given RAGM,*

1. *(C1) and (P) are together equivalent to the single rule*

   (C1P) *If $\neg\alpha \notin B(\mathbb{E} * \beta)$ then $B(\mathbb{E} * \alpha * \beta) = B(\mathbb{E} * (\alpha \wedge \beta))$*

2. *(C2) and (D) are together equivalent to the single rule*

   (C2D) *If $\alpha \leftrightsquigarrow_{\mathbb{E}} \beta$ then $B(\mathbb{E} * \alpha * \beta) = B(\mathbb{E} * \beta)$.*

Both (C1P) and (C2D) provide conditions for the reduction of the two-step revision sequence $\mathbb{E} * \alpha * \beta$ to a single-step revision. (C1P) reduces it to an $(\alpha \wedge \beta)$-revision when α is consistent with a β-revision. (C2D) reduces it to a β-revision, ignoring α completely, when α and β counteract with respect to $\mathbb{E}$. Now, it follows from RAGM that the consequent of (C1P) also obtains when $\neg\beta \notin B(\mathbb{E} * \alpha)$. Putting this together we get a most succinct characterisation of restrained revision.

Proposition 8 *Restrained revision is the unique operator which satisfies RAGM and:*

$$B(\mathbb{E} * \alpha * \beta) = \begin{cases} B(\mathbb{E} * \beta) & \text{if } \alpha \leftrightsquigarrow_{\mathbb{E}} \beta \\ B(\mathbb{E} * (\alpha \wedge \beta)) & \text{otherwise.} \end{cases}$$

We remark that if we were to replace "$\alpha \leftrightsquigarrow_{\mathbb{E}} \beta$" in the first clause above by the stronger "α and β are logically inconsistent", we would obtain instead the characterisation of lexicographic revision given in [12].

5 Restrained Revision as a Composite Operator

As we saw in Section 3, Boutilier's natural revision operator – let us denote it in this section by $\oplus$ – is vulnerable to damaging counterexamples such as the red bird Example 1, and fails to satisfy the very reasonable postulate (P). Although a new input α is accepted in the very next epistemic state $\mathbb{E} \oplus \alpha$, $\oplus$ does not in any way provide for the preservation of α after subsequent revisions. As Hans Rott [14, p. 128] describes it, "[t]he most recent input sentence is always embraced without reservation, the last but one input sentence, however, is treated with utter disrespect". Thus, there seem to be convincing reasons to reject $\oplus$ as a viable operator for performing iterated revision. However, the literature on epistemic state change constantly reminds us that keeping changes *minimal* should be a major concern, and when judged from a purely *minimal change* viewpoint, it is clear that $\oplus$ can't be beat! How can we find our way out of this apparent quandary? In this section we show that the use of $\oplus$ can be retained, *provided* its application is preceded by an *intermediate* operation in which, rather than revising $\mathbb{E}$ *by* new input α, essentially α is revised *by* $\mathbb{E}$.

Given an epistemic state $\mathbb{E}$ and sentence α, let us denote by $\mathbb{E} \triangleleft \alpha$ the result of this intermediate operation. $\mathbb{E} \triangleleft \alpha$ is an epistemic state. The idea is that when forming $\mathbb{E} \triangleleft \alpha$, the information in $\mathbb{E}$ should be maintained. That is, the total preorder $\preceq_{\mathbb{E} \triangleleft \alpha}$ should satisfy

$$v \preceq_{\mathbb{E}} w \text{ implies } v \preceq_{\mathbb{E} \triangleleft \alpha} w. \tag{1}$$

But rather than leaving behind α *entirely* in favour of $\mathbb{E}$, as much of the informational content of α should be preserved in $\mathbb{E} \triangleleft \alpha$ as possible. This is formalised by saying that for any $v \in [\alpha], w \in [\neg\alpha]$, we should take $v \prec_{\mathbb{E} \triangleleft \alpha} w$ as long as this does not conflict with (1) above. It is this second requirement which will guarantee α enough of a "presence" in the revised epistemic state $\mathbb{E} * \alpha$ to help it survive subsequent revisions and allow (P) to be captured. Taken together, the above two requirements are enough to specify $\preceq_{\mathbb{E} \triangleleft \alpha}$ uniquely:

$$v \preceq_{\mathbb{E} \triangleleft \alpha} w \text{ iff } \begin{cases} v \preceq_{\mathbb{E}} w \text{ if } v \in [\alpha] \text{ or } w \in [\neg\alpha], \\ v \prec_{\mathbb{E}} w, \text{ otherwise.} \end{cases} \tag{2}$$

Thus, $\preceq_{\mathbb{E} \triangleleft \alpha}$ is just the lexicographic refinement of $\preceq_{\mathbb{E}}$ by the "two-level" total preorder $\preceq_\alpha$ defined by $v \preceq_\alpha w$ iff $v \in [\alpha]$ or $w \in [\neg\alpha]$. This "backwards revision" operator is not new. In fact it has already been studied by Papini in [13], who proved several properties of it.[5] In particular we do not necessarily have $\alpha \in B(\mathbb{E} \triangleleft \alpha)$ (this will hold only if $\neg\alpha \notin B(E)$), and so $\triangleleft$ does not satisfy RAGM.

Given $\triangleleft$, we can define the composite revision operator $*_\triangleleft$ by setting

$$\mathbb{E} *_\triangleleft \alpha = (\mathbb{E} \triangleleft \alpha) \oplus \alpha \tag{3}$$

This, of course, is reminiscent of the Levi Identity [8], used in AGM theory as a recipe for reducing the operation of revision on *knowledge bases* to a composite operation consisting of contraction plus expansion. In (3), $\oplus$ is playing the role of expansion. The operator $*_\triangleleft$ *does* satisfy RAGM. In fact, as can easily be seen by comparing (2) above with condition (R) at the start of Section 4, $*_\triangleleft$ coincides with restrained revision.

Proposition 9 *Let $*_{\mathrm{R}}$ denote the restrained revision operator. Then $*_{\mathrm{R}} = *_\triangleleft$.*

Thus we have proved that restrained revision can be viewed as a *combination* of two existing operators.

[5] It can also be viewed as just a "backwards" version of Nayak's lexicographic revision operator.

6 Conclusion

We have shown that the Darwiche-Pearl arguments in favour of their framework, taken to their logical conclusion, lead to the acceptance of the admissible revision operators as a class worthy of study. The restrained revision operator, in particular, exhibits quite desirable properties. Besides taking the place of natural revision as the operator adhering most closely to the principle of minimal change, its satisfaction of the properties (O) and (Q) shows that it is not in the business of the unnecessary removal of previously obtained information.

For future work we would like to explore more thoroughly the whole class of admissible revision operators. In this paper we saw that restrained revision and lexicographic revision lie at opposite ends of the spectrum of admissible operators. They represent respectively the most conservative and the least conservative admissible operators. A natural question is whether there exists an axiomatisable class of admissible operators which represents the "middle ground". Perhaps one clue for finding such a class can be found in the counteracts relation $\leftrightsquigarrow_{\mathbb{E}}$ which can be derived from an epistemic state $\mathbb{E}$. As we said, this relation depends only on the preorder $\preceq_{\mathbb{E}}$ associated to $\mathbb{E}$. In fact, given *any* total preorder $\preceq$ over V we can define the relation $\leftrightsquigarrow_{\preceq}$ by

$$\alpha \leftrightsquigarrow_{\preceq} \beta \text{ iff } \min(\alpha, \preceq) \subseteq [\neg\beta] \text{ and } \min(\beta, \preceq) \subseteq [\neg\alpha].$$

Then clearly $\leftrightsquigarrow_{\mathbb{E}} = \leftrightsquigarrow_{\preceq_{\mathbb{E}}}$. Furthermore if $\preceq$ is the full relation $V \times V$ then $\leftrightsquigarrow_{\preceq}$ reduces to the relation of logical inconsistency. A counteracts relation *stronger* than $\leftrightsquigarrow_{\mathbb{E}}$, but still *weaker* than logical inconsistency can be found by setting $\leftrightsquigarrow = \leftrightsquigarrow_{\preceq'}$, where $\preceq'$ lies somewhere *in between* $\preceq_{\mathbb{E}}$ and $V \times V$. Hence one avenue worth exploring might be to assume that from each epistemic state $\mathbb{E}$ we can extract not one but *two* preorders $\preceq_{\mathbb{E}}$ and $\preceq'_{\mathbb{E}}$ such that $\preceq_{\mathbb{E}} \subseteq \preceq'_{\mathbb{E}}$. Then, instead of only requiring $\alpha \leftrightsquigarrow_{\mathbb{E}} \beta$ to deduce $\neg\alpha \in B(\mathbb{E} * \alpha * \beta)$, as is done with restrained revision (the postulate (D)), we could require the stronger condition $\alpha \leftrightsquigarrow_{\preceq'_{\mathbb{E}}} \beta$ for this to hold. We are currently experimenting with strategies for using the second preorder to *guide* the manipulation of $\preceq_{\mathbb{E}}$ to enable this property to be satisfied.

Acknowledgement

The first author wishes to thank Aditya Ghose for making it possible to enjoy a great working environment in Wollongong, and also for some interesting comments on this work. National ICT Australia is funded by the Australia Government's Department of Communications, Information and Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Centre of Excellence program. It is supported by its members the Australian National University, University of NSW, ACT Government, NSW Government and affiliate partner University of Sydney.

References

[1] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] Carlos Areces and Veronica Becher. Iterable AGM functions. In *Frontiers of belief revision*, pages 261–277. Kluwer, Dordrecht, 2001.

[3] Richard Booth. On the logic of iterated non-prioritised revision. In *Conditionals, Information and Inference – Selected papers from the Workshop on Conditionals, Information and Inference, 2002*, volume 3301 of *LNAI*, pages 86–107. Springer-Verlag, Berlin, 2005.

[4] Craig Boutilier. Revision sequences and nested conditionals. In R. Bajcsy, editor, *IJCAI-93. Proceedings of the 13th International Joint Conference on Artificial Intelligence held in Chambery, France, August 28 to September 3, 1993*, volume 1, pages 519–525, San Mateo, CA, 1993. Morgan Kaufmann.

[5] Craig Boutilier. Iterated revision and minimal changes of conditional beliefs. *Journal of Philosophical Logic*, 25(3):263–305, 1996.

[6] Adnan Darwiche and Judea Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89:1–29, 1997.

[7] Michael Freund and Daniel Lehmann. Belief revision and rational inference. Technical Report TR 94-16, The Leibniz Centre for Research in Computer Science, Institute of Computer Science, Hebrew University of Jerusalem, 1994.

[8] Peter Gärdenfors. *Knowledge in Flux : Modeling the Dynamics of Epistemic States*. The MIT Press, Cambridge, Massachusetts, 1988.

[9] Yi Jin and Michael Thielscher. Iterated belief revision, revised. In *Proceedings of IJCAI 2005*. To appear, 2005.

[10] Daniel Lehmann, Menachem Magidor, and Karl Schlechta. Distance semantics for belief revision. *Journal of Symbolic Logic*, 66:295–317, 2001.

[11] Abhaya C. Nayak. Iterated belief change based on epistemic entrenchment. *Erkenntnis*, 41:353–390, 1994.

[12] Abhaya C. Nayak, Maurice Pagnucco, and Pavlos Peppas. Dynamic belief change operators. *Artificial Intelligence*, 146:193–228, 2003.

[13] Odile Papini. Iterated revision operations stemming from the history of an agent's observations. In *Frontiers of belief revision*, pages 281–303. Kluwer, Dordrecht, 2001.

[14] Hans Rott. Coherence and conservatism in the dynamics of belief II: Iterated belief change without dispositional coherence. *Journal of Logic and Computation*, 13(1):111–145, 2003.

[15] Karl Schlechta, Daniel J. Lehmann, and Menachem Magidor. Distance semantics for belief revision. In Yoav Shoham, editor, *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge*, pages 137–145. Morgan Kaufmann, 1996.

[16] Krister Segerberg. Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic*, 39:287–306, 1998.

[17] Wolfgang Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In William L. Harper and Brian Skyrms, editors, *Causation in Decision: Belief, Change and Statistics: Proceedings of the Irvine Conference on Probability and Causation: Volume II*, volume 42 of *The University of Western Ontario Series in Philosophy of Science*, pages 105–134, Dordrecht, 1988. Kluwer Academic Publishers.